Computational and Structural Biotechnology Journal 20 (2022) 3567–3580

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Pan-cancer analyses of synonymous mutations based on tissue-specific codon optimality

Xia Ran [a,b], Jinyuan Xiao [c], Fang Cheng [c], Tao Wang [d], Huajing Teng [e], Zhongsheng Sun [a,b,c,*]

[a] Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China
[b] CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou 325000, China
[d] Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Kaifu District, Changsha, Hunan 410078, China
[e] Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Radiation Oncology, Peking University Cancer Hospital & Institute, Beijing, China

ARTICLE INFO

Article history:
Received 10 March 2022
Received in revised form 22 June 2022
Accepted 3 July 2022
Available online 6 July 2022

Keywords:
Synonymous mutations
Codon optimality
Cell cycle
DNA damage repair deficiency
Cancer

ABSTRACT

Codon optimality has been demonstrated to be an important determinant of mRNA stability and expression levels in multiple model organisms and human cell lines. However, tissue-specific codon optimality has not been developed to investigate how codon optimality is usually perturbed by somatic synonymous mutations in human cancers. Here, we determined tissue-specific codon optimality in 29 human tissues based on mRNA expression data from the Genotype-Tissue Expression project. We found that optimal codons were associated with differentiation, whereas non-optimal codons were correlated with proliferation. Furthermore, codons biased toward differentiation displayed greater tissue specificity in codon optimality, and the tissue specificity of codon optimality was primarily present in amino acids with high degeneracy of the genetic code. By applying tissue-specific codon optimality to somatic synonymous mutations in 8532 tumor samples across 24 cancer types and to those in 416 normal cells across six human tissues, we found that synonymous mutations frequently increased optimal codons in tumor cells and cancer-related genes (e.g., genes involved in cell cycle). Furthermore, an elevated frequency of optimal codon gain was found to promote tumor cell proliferation in three cancer types characterized by DNA damage repair deficiency and could act as a prognostic biomarker for patients with triple-negative breast cancer. In summary, this study profiled tissue-specific codon optimality in human tissues, revealed alterations in codon optimality caused by synonymous mutations in human cancers, and highlighted the non-negligible role of optimal codon gain in tumorigenesis and therapeutics.

1. Introduction

Cancer is a leading cause of morbidity and mortality worldwide, with 19.3 million new cases and almost 10.0 million deaths in 2020 [1]. Among the diverse cancer types, breast cancer is the most commonly diagnosed cancer [1], and triple-negative breast cancer (TNBC) is a highly aggressive subtype with poorer clinical outcome and greater metastatic potential [2]. To unveil the mystery of cancer occurrence, progression, and outcome, genome sequencing projects covering high-volume cancer samples have been carried out, and a wide variety of genomic mutations have been identified [3]. Despite being the second most common type of point mutation [4], synonymous mutations (i.e., substitutions between synonymous codons with no change in the amino acid composition of the encoded proteins) are generally considered silent in cancer due to the dogma that amino acid sequences determine protein structure and function. Same amino acids as they encode, the synonymous codons are unevenly used in the transcriptome. Indeed, codon usage bias (or codon bias) widely exists in multiple domains of life [5–7] and is thought to be shaped to some extent by codon optimality, which refers to the non-uniform decoding rate of the 61 amino-acid encoding codons by the ribosome due to the variability of tRNA concentrations and the stochastic nature of ribosome decoding [8]. In recent years, codon optimality has been revealed to be an important determinant of mRNA stability in multiple model organisms and human cell lines [9–13]. In line with this, codon usage has been shown to be an important determinant of mRNA expression levels [14,15]. Furthermore, it has been reported in the past few years that codon usage exhibits cancer-specific patterns [16–20]. However, it remains largely unexplored how codon

* Corresponding author at: Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China.
E-mail address: sunzs@biols.ac.cn (Z. Sun).

https://doi.org/10.1016/j.csbj.2022.07.005

optimality is usually perturbed by synonymous mutations in human cancers.

Several metrics have been developed to measure codon optimality based on either or both of tRNA levels and codon usage [9,21–23]. In this context, Supek et al. [24] investigated whether synonymous mutations increased optimal codons in oncogenes and enhanced oncogene translation efficiency or accuracy, yet no evidence was found based on the genomic tRNA gene copy number. By applying codon stabilization coefficient (CSC) scores from the human embryonic kidney 293T cell line (HEK293T) to germline synonymous mutations in healthy individuals, Dhindsa et al. [25] found that DNA damage-response genes and cell-cycle regulated genes were particularly intolerant to synonymous mutations. Notably, both tRNA abundance [26,27] and codon usage [28,29] have been reported to exhibit differences between human tissues. However, tissue-specific codon optimality has not been developed to investigate the impact of synonymous mutations on codon optimality in human cancers. Additionally, although the calculation of CSC scores requires mRNA half-life data that are not available in diverse human tissues, three other existing metrics, including the tRNA adaption index (tAI) [21], C_{opt} (i.e., codon optimality) [22], and the normalized translation efficiency (nTE) [23], can be employed to generate tissue-specific codon optimality based on the extant tRNA abundance [30] and mRNA expression data [31] in human tissues. Still, it remains unclear which metric measures codon optimality more accurately in *Homo sapiens*.

In this study, to investigate the impact of synonymous mutations on codon optimality in human cancers, we first evaluated the three metrics (tAI, C_{opt}, and nTE) in HEK293T cells. We found that C_{opt} outperformed tAI and nTE in measuring codon optimality in human tissues and that our modified version of C_{opt}, called rate ratio (RR) score, measured codon optimality more accurately than C_{opt} in human tissues. The RR score was then applied to generate tissue-specific codon optimality for human tissues based on expression data across 29 human tissues from the Genotype-Tissue Expression (GTEx) project [31]. Interestingly, optimal codons were found to be associated with differentiation, while non-optimal codons were correlated with proliferation; codons biased toward differentiation displayed greater tissue specificity in codon optimality. Additionally, the tissue specificity of codon optimality was primarily observed for amino acids with high degeneracy of the genetic code. By applying tissue-specific codon optimality to somatic synonymous mutations in 8532 tumors samples across 24 cancer types from The Cancer Genome Atlas (TCGA) and to those in 416 normal cells from 234 individuals across six human tissues from a Database of Somatic Mutations in Normal Cells (DSMNC) [32], we found that synonymous mutations frequently increased optimal codons in tumor cells and cancer-related genes (e.g., genes involved in cell cycle). Further analyses revealed that an elevated frequency of optimal codon gain promoted tumor cell proliferation in three cancer types characterized by DNA damage repair (DDR) deficiency. Additionally, an elevated frequency of optimal codon gain was found to correlate with better survival in patients with TNBC. These findings may provide insights into how synonymous mutations contribute to tumor progression and outcome by altering codon optimality, and may help to uncover the mystery of cancer occurrence, progression and outcome.

2. Materials and methods

2.1. Calculation of tAI, nTE, C_{opt}, RR score, and codon-specific ribosome density in HEK293T cell line

tAI was calculated based on tRNA levels and Crick's wobble rules to reflect how efficiently tRNA was used by the ribosome [21]. The copy number of tRNA genes (hg38) was obtained from the Genomic tRNA Database (GtRNAdb) [33], and tRNA abundance quantified by demethylase tRNA-seq by Zheng et al. [34] was downloaded from a public dataset in Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66550). tAI was calculated based on the copy number of tRNA genes and square-root-normalized tRNA abundance, respectively, using the R package *tAI* [21].

nTE was calculated by normalizing the tAI by codon usage [23]. RNA-seq data (adaptor trimmed) of HEK293T cells were downloaded from a public dataset in GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113952). Trimmomatic v0.39 [35] was used for quality trimming and RSEM v1.2.30 [36] was employed for transcript-level quantification based on reads mapped against the reference genome (hg38) using STAR v2.7.0c [37]. The coding sequence (CDS) profile of the human genome (hg38) was obtained from the Ensemble database (ftp://ftp.ensembl.org/pub/release-100/fasta/homo_sapiens/cds/Homo_sapiens.GRCh38.cds.all.fa.gz), and the number of occurrences for each codon in an open reading frame (ORF) was generated for 111,267 coding transcripts by a Perl program named 'codonM' from the R package *tAI* [21]. Codon usage was calculated based on codon occurrence and square-root-normalized transcript abundance, and nTE was then calculated by normalizing the tAI by codon usage, as previously described [23], based on the above two types of tAI, respectively. Notably, the first codon in every sequence was ignored when counting the number of occurrences of each codon in a transcript since it is always a methionine codon (even if it is not coded by the canonical ATG). Accordingly, tRNA copy number or abundance of iMetCAT was not included when calculating the tAI of ATG.

C_{opt} was calculated by comparing codon usage in highly and lowly expressed genes based on the assumption that highly expressed genes are codon-optimized [22]. Based on the above codon occurrence data and mRNA expression data of HEK293T cells, codon occurrence was counted for each codon and each amino acid family in highly (transcripts per million [TPM] $\geq$ 100) and lowly (5 > TPM $\geq$ 1) expressed genes, respectively. C_{opt} was calculated as the odds ratio of codon occurrence in highly versus lowly expressed genes using Fisher's exact test, while RR score, our modified version of C_{opt}, was calculated as the RR of codon occurrence in highly versus lowly expressed genes using the RR test from the R package *rateratio.test*.

Ribosome profiling data of HEK293T cells were obtained from a previous study by Ingolia et al. [38]. Adapter and quality trimming were performed using Cutadapt v2.10 [39] and Trimmomatic v0.39 [35], respectively, according to Liu et al. [40]. The trimmed reads were then aligned to rRNAs using Bowtie v1.1.2 [41] with default parameters to avoid rRNA contamination. The unmapped reads were then mapped against coding sequences using STAR v2.7.0c [37] with the following parameters '--outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate --quantMode TranscriptomeSAM GeneCounts --outFilterMultimapNmax 1 --outFilterMatchNmin 16 --alignEndsType EndToEnd'. RiboMiner [42] was used for periodicity checking and calculation of the ribosome density at each position for each transcript. Codon-specific ribosome density was then calculated according to Weinberg et al. [43].

2.2. Evaluation of metrics measuring codon optimality in the HEK293T cell line

To determine which metric measured codon optimality more accurately in *Homo sapiens*, an evaluation was performed by assessing the Pearson correlation of each of the four metrics (tAI, C_{opt}, nTE, and RR score) with codon-specific ribosome density

and CSC score (i.e., the Pearson correlation coefficient between codon usage and mRNA stability) in HEK293T cells. CSC scores of HEK293T were obtained from a recent study by Wu et al. [11], which included CSC scores derived from endogenous (reflecting mRNA decay regulation from codon composition) or ORFome (reflecting mRNA decay regulation from other regulatory information, such as untranslated regions) mRNA. The relationship between RR score and the other three metrics was also evaluated using the Pearson correlation analyses. Theoretically, a negative correlation exists between codon optimality and ribosome density, while a positive correlation exists between codon optimality indices. Therefore, the correlations between CSC score and the four metrics were examined using one-sided Pearson correlation analysis with the option of alternative= 'greater', which was also employed to evaluate the correlations between RR score and other three metrics of codon optimality. The correlations between the four metrics and ribosome density were evaluated using one-sided Pearson correlation analysis with the option of alternative= 'less'.

2.3. Generation of tissue-specific codon optimality for human tissues

Transcript-level mRNA expression data of human tissues were obtained from the GTEx Portal (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz) and 29 human tissues with expression data available in >10 samples were used, including adipose tissue, adrenal gland, bladder, blood, blood vessel, brain, breast, cervix, colon, esophagus, heart, kidney, liver, lung, muscle, nerve, ovary, pancreas, pituitary, prostate, salivary gland, skin, small intestine, spleen, stomach, testis, thyroid, uterus, and vagina tissues (Table S1). Given the large number of samples in most tissues, quality control was performed according to a previous study [44] as follows: i) a standard sample was defined for each tissue as the median value of TPM of all samples from the same tissue; ii) the Spearman correlation coefficient of TPM was calculated between each sample and its corresponding standard sample; iii) samples with a correlation of >0.8 were defined as qualified.

The median expression of a transcript in all qualified samples from the same tissue was taken as the transcript abundance in the tissue, and transcripts that were among the 111,267 coding transcripts were then used to identify the highly (TPM $\geq$ 100) and lowly (5 > TPM $\geq$ 1) expressed genes in each human tissue. Tissue-specific codon optimality was then calculated for each of the 29 human tissues as aforementioned. For each human tissue, codons were defined as optimal if they were significantly overrepresented in highly expressed genes compared to lowly expressed genes (i.e., RR score > 1 and $p < 0.05$), and non-optimal otherwise.

2.4. Characterization of tissue-specific codon optimality in human tissues

To characterize tissue-specific codon optimality in human tissues, hierarchical clustering was first performed based on RR scores in 29 human tissues in R with the options of a 'Euclidean' distance and a 'ward.D' linkage. The association of codon optimality with proliferation and differentiation was analyzed by evaluating the Pearson correlation between the mean value of RR scores in 29 human tissues and the log2 ratio of codon usage in proliferation- versus differentiation-related genes. Codon usage in proliferation- and differentiation-related genes was obtained from Gingold et al. [45]. The standard variations of distribution of RR scores were then examined to identify the codons whose codon optimality exhibited larger variations among human tissues. Hierarchical clustering was also performed, as described above, based on the optimal and non-optimal codons in 29 human tissues. The number of optimal codons was counted for each amino acid in each of the human tissues, and hierarchical clustering was performed as described above to characterize the distribution of tissue-specific optimal codons at the amino acid level.

2.5. Identification of somatic synonymous mutations altering codon optimality in tumor samples from TCGA

Somatic mutation data were obtained from TCGA public access portal for 24 cancer types whose primary sites were within the 29 human tissues with tissue-specific codon optimality data available, including Adrenocortical Carcinoma (ACC) and Pheochromocytoma and Paraganglioma (PCPG) from the adrenal gland, Bladder Urothelial Carcinoma (BLCA) from the bladder, Glioblastoma Multiforme (GBM) and Brain Lower grade Glioma (LGG) from the brain, Breast Invasive Carcinoma (BRCA) from the breast, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) from the cervix, Colon Adenocarcinoma (COAD) from the colon, Esophageal Carcinoma (ESCA) from the esophagus, Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), and Kidney Renal Papillary Cell Carcinoma (KIRP) from the kidney, Liver Hepatocellular Carcinoma (LIHC) from the liver, Lund Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) from the lung, Ovarian Serous Cystadenocarcinoma (OV) from the Ovary, Pancreatic Adenocarcinoma (PAAD) from the pancreas, Prostate Adenocarcinoma (PRAD) from the prostate, Skin Cutaneous Melanoma (SKCM) from the skin, Stomach Adenocarcinoma (STAD) from the Stomach, Testicular Germ Cell Tumors (TGCT) from the testis, Thyroid Carcinoma (THCA) from the thyroid, Uterine Corpus Endometrial Carcinoma (UCEC) and Uterine Carcinosarcoma (UCS) from the uterus (Table S2).

False positive somatic mutations were filtered as previously described [46]. Based on the tissue-specific optimal and non-optimal codons determined above, the resulting somatic synonymous mutations were classified as optimal codon gain if the mutation substituted a non-optimal codon with an optimal codon, optimal codon loss if the mutation replaced an optimal codon with a non-optimal codon, or no change otherwise.

2.6. Characterization of synonymous mutations altering codon optimality in tumor samples from TCGA

Given the existence of tissue-specifically optimal codons, the type of alterations in codon optimality caused by synonymous mutations was counted for each mutation, and the function *oncoprint* from the R package *ComplexHeatmap* was used to visualize the synonymous mutations causing two or three types of codon optimality changes.

Variant allele frequency (VAF) was calculated for each synonymous mutation as the percentage of variant supporting reads in the total depth of the mutated position, and the number of synonymous mutations was calculated for the above three types of synonymous mutations in each VAF subgroup (i.e., 0%-10%, 10%-20%, 20%-30%, 30%-40%, 40%-50%, 50%-60%, 60%-70%, 70%-80%, 80%-90%, 90%-100%). The relative location was calculated for each synonymous mutation as the percentage of mutated codon position in the total length of amino acids of the corresponding transcripts, and the number of synonymous mutations was calculated for the above three types of synonymous mutations in each location subgroup (i.e., 0%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–50%, 50%–60%, 60%–70%, 70%–80%, 80%–90%, 90%–100%). The RR test was used to compare optimal codon gain and optimal codon loss or optimal codon gain and no change in the VAF subgroup of 0%–10%, as well as to compare optimal codon gain and optimal codon loss in the location subgroups of 0%–20% and 60%–80%.

For optimal codon gain or loss, the occurrence of codon change was counted for each cancer type and normalized by the maximum

occurrence in the corresponding cancer type. Hierarchical clustering was performed as aforementioned based on the top three most frequent optimal codon gains or losses ranked by mutation occurrence in each cancer type.

2.7. Validation of the impact of somatic synonymous mutations on codon optimality in the HeLa cell line

CSC scores for the HeLa cell line were obtained from a recent study by Wu et al. [11]. Changes in CSC scores in the HeLa cell line were examined for somatic synonymous mutations causing optimal codon gains, optimal codon losses, or no change in CESC tumor samples. The Wilcoxon rank-sum test was used to compare the changes in CSC scores between optimal codon gains or losses and no changes. In addition, changes in CSC scores in the HeLa cell line were examined for the top three most frequent optimal codon gains or losses in CESC.

2.8. Identification of the genes in which synonymous mutations frequently increase or decrease optimal codons

The genes in which synonymous mutations frequently increased or decreased optimal codons in tumor cells were identified by comparing the frequency of optimal codon gain or loss (i.e., the fraction of synonymous mutations causing optimal codon gain or loss) in each gene with that in a set of nonessential genes (n = 16,331) obtained from Kumar et al. [47] using the RR test. P-values were adjusted using the 'false discovery rate (FDR)' method.

2.9. Comparing the frequency of optimal codon gain (or loss) between cancer-related genes and non-essential genes

Cancer-related genes, including those involved in apoptosis (n = 140), cell cycle (n = 124), DNA repair (n = 178), immune response (n = 4723), and metabolism (n = 1939), as well as genes which were nonessential for tumor cell growth (n = 16,331), were obtained from Kumar et al. [47]. The frequency of optimal codon gain or loss was computed as the fraction of synonymous mutations causing optimal codon gains or losses in each gene set and then compared between each cancer-related gene set and nonessential gene set using the RR test. The frequency of optimal codon gain or loss was also calculated for each gene, and the Wilcoxon rank-sum test was used to compare the frequency of optimal codon gain between genes related to cell cycle (or DNA repair) and nonessential genes, as well as to compare the frequency of optimal codon loss between metabolism-related genes and nonessential genes.

2.10. Comparing the frequency of optimal codon gain (or loss) between tumor and normal tissues

Somatic mutations in 416 normal cells from 234 individuals across six human tissues (Table S3) were obtained from DSMNC [32], and variant effect predictor (VEP) [48] was used to annotate the codon changes caused by synonymous mutations. The impact of synonymous mutations on codon optimality was analyzed as described above. The frequency of optimal codon gain (or loss) was computed as the fraction of synonymous mutations causing optimal codon gains or losses in each tumor sample or normal cell, and compared between tumor and normal tissues using the Wilcoxon rank-sum test. We also obtained somatic mutations in 1059 normal samples across 28 human tissues (Table S4) from a previous study that detected somatic mutations based on RNA-seq data in GTEx [49], and performed similar analyses. To analyze the influence of mutation rate differences between cancer types on the frequency of optimal codon gain or loss, Pearson correlation analysis was employed to evaluate the correlation between the ranked position of cancer types with respect to the load of synonymous mutations and the ranked position of cancer types with respect to the load or frequency of optimal codon gain or loss.

2.11. Gene ontology (GO) enrichment analysis

WebGestalt [50] was used to identify enriched GO biological processes using default parameters (i.e., the top 10 enriched terms ranked based on FDR).

2.12. Identification of the genes whose mRNA expression levels increase with elevated frequency of optimal codon gain

Genes whose mRNA expression levels increased with an elevated frequency of optimal codon gain were identified as those whose expression levels were positively correlated with the frequency of optimal codon gain in each cancer type. The correlation between the frequency of optimal codon gain and mRNA expression level of each gene was evaluated using the Spearman correlation analysis, and p-values were adjusted using the 'FDR' method.

2.13. Association between the frequency of optimal codon gain and tumor cell proliferation

For each tumor sample in TCGA, proliferation scores were obtained from Thorsson et al. [51]. The association between the frequency of optimal codon gain and two metrics reflecting tumor cell proliferation (i.e., ki-67 mRNA expression level, and proliferation) was evaluated using the Spearman rank correlation analysis for each cancer type. P-values were adjusted using the 'FDR' method.

2.14. Association between the frequency of optimal codon gain and DDR deficiency

To measure DDR deficiency, 71 DNA repair pathway-specific genes were obtained from Knijnenburg et al. [52], which included genes involved in base excision repair (BER, n = 8), direct repair (DR, n = 3), Fanconi anemia (FA, n = 8), homologous recombination (HR, n = 21), mismatch repair (MMR, n = 8), nucleotide excision repair (NER, n = 10), non-homologous end joining (NHEJ, n = 8), and translesion synthesis (TLS, n = 5). Based on somatic mutations in BRCA, STAD, and UCEC, each type of DDR deficiency was calculated as the weighted sum of the mutations in each set of the DDR genes. The weight was 1 for missense mutations; 5 for nonsense, splice site, or frameshift mutations; 10 for gene-level loss of one copy number; and 20 for gene-level loss of two copy numbers. For each sample in TCGA, HR defects were obtained from Thorsson et al. [51]. The Spearman rank correlation analysis was used to evaluate the association between the frequency of optimal codon gain and HR defects or each type of DDR deficiency.

2.15. Association between the frequency of optimal codon gain and patient survival

For patients with BRCA, STAD, and UCEC in TCGA, clinical information, including patient survival data, was obtained from a previous study by TCGA [53]. Based on the immunohistochemistry status of estrogen receptor (ER), erb-b2 receptor tyrosine kinase 2 (ERBB2 or HER2), and progesterone receptor (PR) obtained from cBioPortal, patients with BRCA in TCGA were classified as luminal A (ER- and PR-positive, HER2-negative), HER2+ (ER- and PR-negative, HER2-positive), TNBC (ER-, PR- and HER-negative), or luminal B (ER-positive) according to Goldhirsch et al. [54]. In addition, somatic synonymous mutations and corresponding clinical infor-

mation were obtained for 147 patients with TNBC receiving adjuvant chemotherapy from a recent study by Staaf et al. [55]. Codon changes caused by synonymous mutations were annotated using VEP [48], and the impact of synonymous mutations on codon optimality was analyzed as described above. The distribution of patient survival was examined using the Kaplan-Meier method and was compared using the log-rank test. Hazard ratios and 95% confidence intervals (CIs) were determined using univariate and multivariate Cox proportional hazard regression models.

2.16. Statistical analyses

Unless otherwise specified, all p-values were two-sided, and p-values <0.05 were considered statistically significant. All statistical analyses were performed using R (version 3.5).

3. Results

3.1. RR score is a better measurement of codon optimality in Homo sapiens

To investigate how codon optimality is perturbed by synonymous mutations in human cancers, we first determined which index measured codon optimality more accurately in *Homo sapiens*. This analysis was conducted in the HEK293T cell line by evaluating the association of three indices, including tAI, nTE, and C_{opt}, with codon-specific ribosome density and CSC score. An index would be considered if it was negatively associated with codon-specific ribosome density and positively correlated with CSC score. We first examined the correlation of tAI with codon-specific ribosome density and CSC score. tAI was calculated as previously described [21], based on the copy number of tRNA genes (hg38). A positive correlation was observed between tAI and CSC scores derived from endogenous mRNAs (R = 0.25, p = 0.0248; Fig. S1A). A similar trend was observed between tAI and CSC scores derived from ORFome mRNAs (R = 0.16, p = 0.1044; Fig. S1B), yet significant association was not observed between tAI and codon-specific ribosome density (p = 0.4179; Fig. S1C). We also calculated tAI based on tRNA abundance of HEK293T cells. Although a stronger correlation was observed between tAI and CSC scores derived from endogenous (R = 0.39, p = 0.0009; Fig. S1D) or ORFome mRNAs (R = 0.26, p = 0.0213; Fig. S1E), significant association was still not observed between tAI and codon-specific ribosome density (p = 0.6118; Fig. S1F). We then calculated nTE by normalizing tAI by codon usage, as previously described [23], and analyzed its association with codon-specific ribosome density and CSC score. Consistent with previous observations in yeast [23], a positive relationship was observed between tAI and codon usage (R = 0.38, p = 0.0022; Fig. S2A). Although nTE trended to be negatively correlated with codon-specific ribosome density, the correlation was not statistically significant (R = −0.19, p = 0.0692; Fig. S2B). Additionally, significant correlation was not observed between nTE and CSC scores from either endogenous (R = −0.06, p = 0.6866; Fig. S2C) or ORFome mRNAs (R = −0.03, p = 0.5775; Fig. S2D). We also calculated nTE using the tAI calculated from tRNA abundance. Although a stronger correlation was observed between tAI and codon usage (R = 0.46, p = 0.0002; Fig. S2E), nTE showed no significant correlation with codon-specific ribosome density (R = −0.12, p = 0.1789; Fig. S2F) or CSC scores (endogenous mRNAs: R = −0.05, p = 0.6431; ORFome mRNAs: R = −0.01, p = 0.53; Fig. S2G-H). Next, we evaluated the correlation of C_{opt} with codon-specific ribosome density and CSC score. Interestingly, C_{opt} showed a significantly negative correlation with codon-specific ribosome density (R = −0.22, p = 0.0427; Fig. S3A) and a significantly positive correlation with CSC scores from either endogenous (R = 0.288, p = 0.0122; Fig. S3B) or ORFome mRNAs (R = 0.415, p = 0.00004; Fig. S3C). These data suggest that C_{opt} measures codon optimality more accurately than tAI and nTE in *Homo sapiens*.

In view of the above results, we introduced a modified version of C_{opt} by calculating the RR (i.e., RR score) rather than the odds ratio (i.e., C_{opt}) of codon usage in highly versus lowly expressed genes, and examined its correlation with codon-specific ribosome density and CSC score. As expected, codon-specific ribosome density showed a significantly negative correlation with RR score (R = −0.24, p = 0.0294; Fig. 1A), which was slightly stronger than its correlation with C_{opt}. Furthermore, CSC scores from either endogenous (R = 0.54, p = 3.65e−06; Fig. 1B) or ORFome (R = 0.68, p = 6.75e−10; Fig. 1C) mRNAs exhibited a much stronger correlation with RR score than with C_{opt}. These data suggest that RR score outperforms C_{opt} in measuring codon optimality in *Homo sapiens*. Further examination of the associations between RR score and the other three metrics revealed that RR score showed a weakly yet significantly positive relationship with tAI (R = 0.25, p = 0.0264; Fig. 1D), no significant correlation with nTE (R = 0.16, p = 0.1112; Fig. 1E), and a significantly positive correlation with C_{opt} (R = 0.75, p = 1.86e−12; Fig. 1F). Overall, these data suggest that RR score is a better measurement of codon optimality in *Homo sapiens*.

3.2. Tissue-specific codon optimality in 29 human tissues

Next, RR score was used to calculate tissue-specific codon optimality for human tissues based on mRNA expression data from 15,533 qualified (see Methods; Table S1) samples across 29 human tissues from GTEx [31]. Similar to the definition of optimal codons in C_{opt} [22], codons were defined as optimal if they were significantly overrepresented in highly expressed genes (i.e., RR score > 1 and p < 0.05 by RR test) and non-optimal otherwise.

To characterize tissue-specific codon optimality in human tissues, we first performed hierarchical clustering based on RR scores in 29 human tissues (Table S5). As a result, 24 of the 61 amino acid encoding codons had a widespread RR score of >1, while another 26 codons' RR scores were generally <1 (Fig. 2A). Interestingly, 22 of the 24 codons ended with a C or G, whereas 23 of the 26 codons ended with an A or T. Notably, previous studies revealed two types of codons related to proliferation and differentiation, respectively [45,56]. Thus, we examined the association of codon optimality with proliferation and differentiation. Intriguingly, a significantly negative correlation was observed between RR score and log2ratio of codon usage in proliferation- and differentiation-related genes (R = −0.75, p = 5e−12; Fig. 2B), which pointed to the association of optimal and non-optimal codons with differentiation and proliferation, respectively. To investigate the tissue specificity of codon optimality, we examined the standard deviations of RR scores in the 29 human tissues for all the 61 amino acid encoding codons. Interestingly, nine codons were clearly distinguished from the rest by relatively large deviations in RR scores, and seven out of the nine codons had a mean RR score of >1 (Fig. 2C). This suggests that codons biased toward differentiation exhibit greater tissue specificity in codon optimality. Additionally, five of the nine codons ended with CG (i.e., CCG, GCG, TCG and ACG) or with GC (i.e., CGC; Fig. 2D), which were exactly the top five codons biased toward differentiation (Fig. 2E). These data suggest that differentiation may be a potential driver of tissue specificity in codon optimality.

We also performed hierarchical clustering based on the optimal and non-optimal codons in 29 human tissues (Table S6). As a result, 24 codons were optimal in all (n = 19) or most (n = 5) of the tissues, whereas the remaining 37 codons were optimal in few (n = 7) or none (n = 30) of the tissues (Fig. S4). That is to say, there are 12 codons being tissue-specifically optimal in the

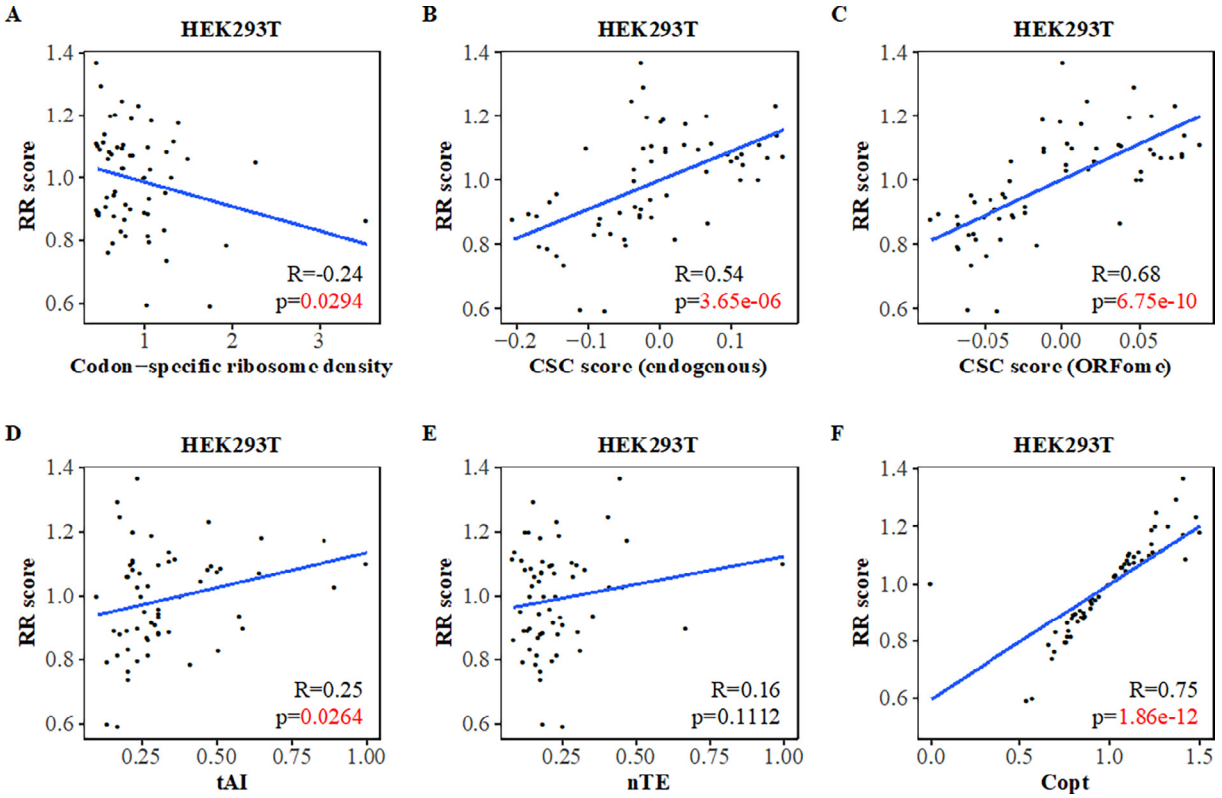

Fig. 1. RR score is a better measurement of codon optimality in Homo sapiens (A-F) The association of RR score with codon-specific ribosome density (A), CSC score based on endogenous mRNAs (B), CSC score based on ORFome (C), tAI (D), nTE (E), and C_{opt} (F) in HEK293T cells.

29 human tissues, including the 4 codons ending with CG. In line with aforementioned results, two-thirds (8/12) of the tissue-specifically optimal codons were biased toward differentiation (Fig. 2F). We also evaluated tissue-specific optimal codons at the amino acid level. Interestingly, the number of optimal codons in nine amino acids (i.e., Leu, Ser, Arg, Ala, Gly, Val, Thr, Pro, and Gln) varied across 29 human tissues, and eight of the nine amino acids were encoded by four- (i.e., Ala, Gly, Val, Thr and Pro) or six-fold (i.e., Leu, Ser and Arg) degenerate codons (Fig. 2G); that is to say, each of the eight amino acids with high degeneracy of the genetic code (i.e., four- or six-fold, relative to two- or three-fold) had at least one codon being tissue-specifically optimal in the 29 human tissues. These data suggest that the tissue specificity of codon optimality is primarily present in amino acids with high degeneracy of the genetic code.

3.3. Identification of somatic synonymous mutations altering codon optimality in 24 cancer types

Next, we investigated the impact of somatic synonymous mutations on codon optimality in human cancers. A total of 458,611 nonredundant somatic synonymous mutations were obtained from 8,532 tumor samples across 24 cancer types (originating from 18 human tissues where codon optimality data were available; Table S4) in TCGA. Based on tissue-specific optimal and non-optimal codons, a synonymous mutation was classified as an optimal codon gain if a non-optimal codon was substituted with an optimal codon, an optimal codon loss if an optimal codon was replaced with a non-optimal codon, or no change (i.e., remained optimal or non-optimal; Fig. 3A). As a result, 99.34% (455,577/458,611) of the synonymous mutations consistently resulted in optimal codon gain (54,150; 12%), optimal codon loss (281,332; 62%), or no change (120,095; 26%) in corresponding cancer types (Fig. 3B and C). The remaining 3034 synonymous mutations led to two (n = 3031; Fig. 3D) or three (n = 3; Fig. 3E) different types of alterations in codon optimality, as these synonymous mutations were mutated in at least two cancer types and the involved codons were tissue-specifically optimal in human tissues. For example, a synonymous mutation in *MSH6* (c.957G > A, ACG->ACA) which occurred in five cancer types including UCEC, BLCA, PRAD, GBM, and STAD, caused optimal codon loss in UCEC, BLCA, and PRAD, yet no change in GBM and STAD, as the codon ACA was non-optimal in all the human tissues, while the codon ACG was optimal in only 6 human tissues including Uterus, Bladder, Prostate, Cervix, Thyroid, and Pituitary. For validation, CSC scores of the HeLa cell line were obtained from Wu et al. [11] to examine changes in CSC scores for all the synonymous mutations causing optimal codon gains, optimal codons losses, or no changes in cervical cancer (i.e., CESC). Compared with synonymous mutations causing no changes to codon optimality, those leading to optimal codon gains were significantly more likely to enhance codon optimality ($p < 2e{-}16$, Wilcoxon rank-sum test), while those resulting in optimal codon losses were significantly more likely to attenuate codon optimality ($p < 2e{-}16$, Wilcoxon rank-sum test; Fig. 3F). Similar trends were observed when examining changes in CSC score for the top three most frequent optimal codon gains or losses in CESC (Fig. 3G).

Based on the top three most frequent optimal codon losses in each cancer type, the 24 cancer types were divided into two clusters by hierarchical clustering. One cluster (C1) comprised 10 cancer types, including four types of gynecological cancers (BRCA, CESC, UCEC and UCS), GBM, LUSC, SKCM, ESCA, BLCA, and THCA. The other cluster (C2) comprised 14 cancer types, including five cancer types dominated by adenocarcinoma (LUAD, PAAD, COAD, STAD, and PRAD), three cancer types originating from the kidney (KIRC, KIRP, and KICH), two cancer types originating from the adre-

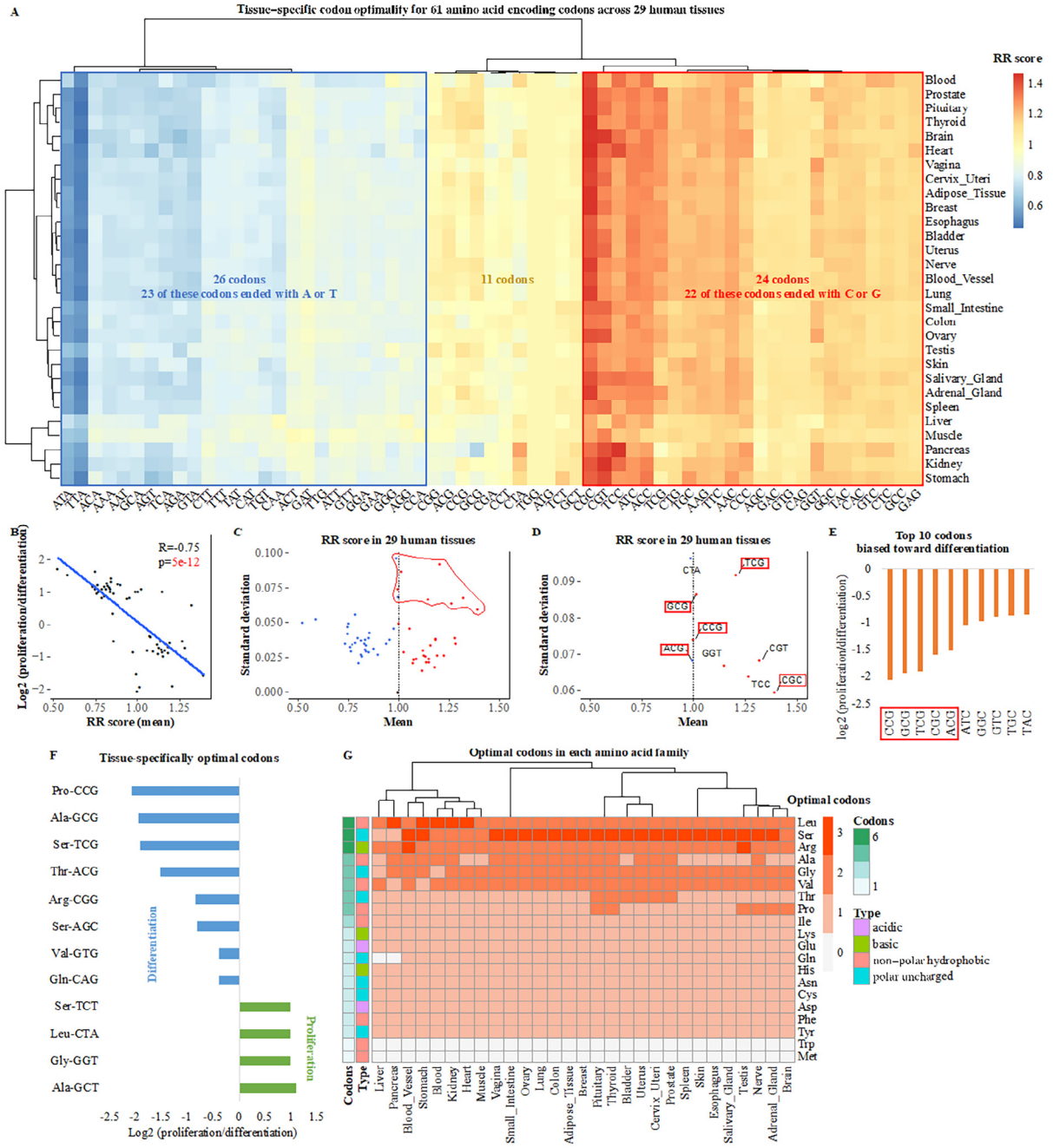

Fig. 2. Tissue-specific codon optimality in 29 human tissues (A) Hierarchical clustering based on RR scores of 61 amino acid encoding codons in 29 human tissues. (B) Association between the mean RR scores in 29 human tissues and the log2 ratio of codon usage in proliferation- versus differentiation-related genes. (C) Mean values and standard deviations of RR scores in 29 human tissues. (D) Nine codons with relatively large deviations in RR scores in 29 human tissues. (E) Top 10 codons biased toward differentiation. (F) Twelve codons being tissue-specifically optimal in 29 human tissues. (G) Tissue-specific optimal codons at the amino acid level.

nal gland (ACC and PCPG), LGG, TGCT, LIHC, and OV. The top three most frequent optimal codon losses in most cancer types were TTC->TTT, ATC->ATT, and GAC->GAT (Fig. 3H). Interestingly, TTC->TTT was the most frequent optimal codon loss in almost all the cancer types in C1, whereas GAC->GAT was the most frequent optimal codon loss in most cancer types in C2 (Fig. 3H). Hierarchical clustering based on the top three most frequent optimal codon gains in each cancer type distinguished three cancer types (LUAD, LUSC, and ACC) from the other 21 cancer types; GGG->GGT and CGG->CGT were the most frequent optimal codon gains in the three cancer types, while AAA->AAG and GAA->GAG were the most frequent optimal codon gains in many of the other 21 cancer types (Fig. 3I).

3.4. Synonymous mutations frequently increase optimal codons in 16 cancer types

To characterize the distribution of optimal codon gains or losses in human cancers, we first analyzed the VAFs of synonymous mutations causing optimal codon gains or losses. Interestingly, optimal codon gains showed higher proportion than optimal codon loss in synonymous mutations with a VAF of <10% ($p < 2e{-}16$, RR test), and this trend was also observed when comparing with no change ($p < 2e{-}16$, RR test; Fig. 4A). This suggests that the increase of optimal codons by somatic synonymous mutations preferentially occurs in subclonal tumor cells. We then analyzed the distribution of optimal codon gains or losses within the CDS.

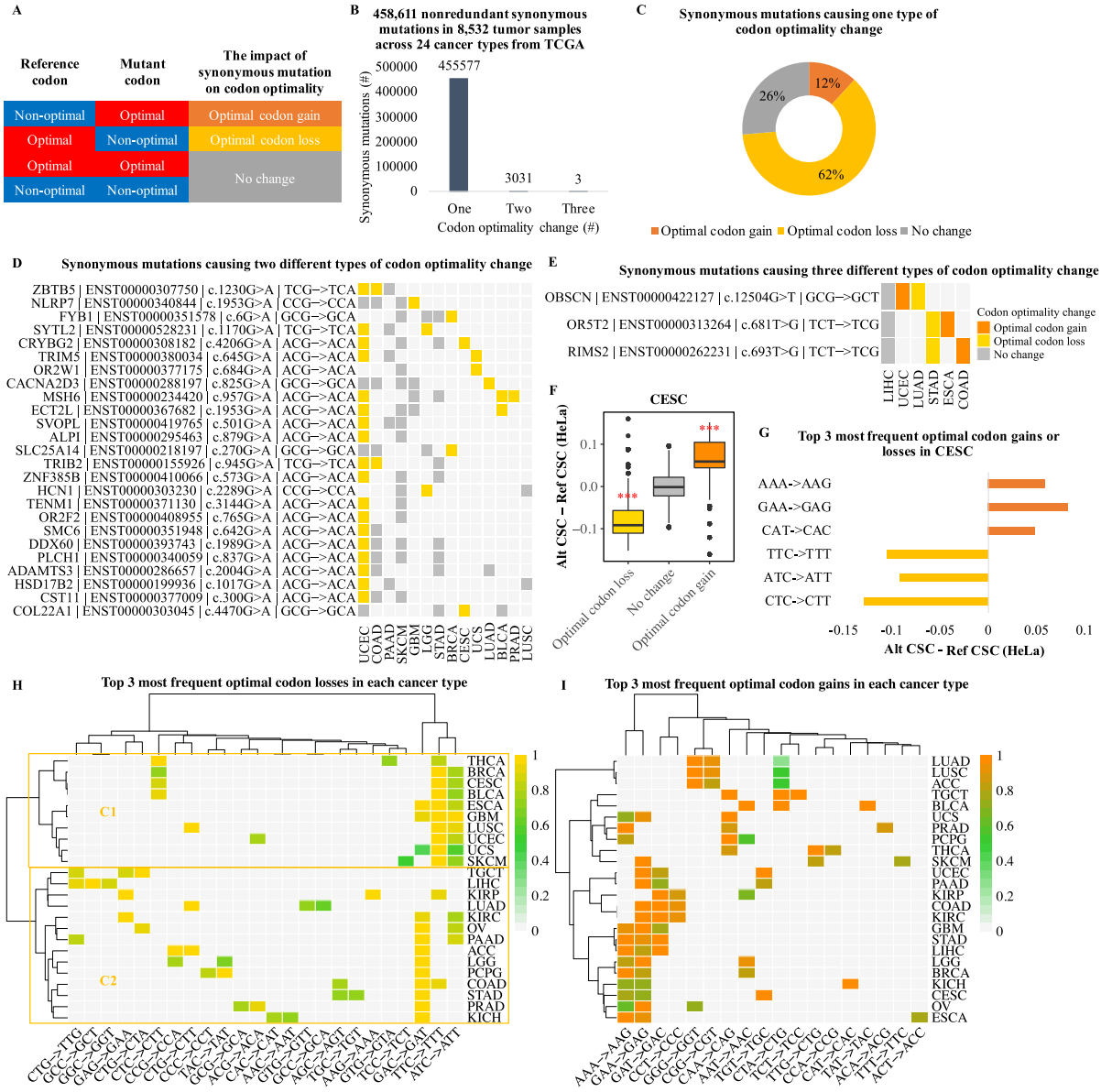

Fig. 3. Identification of somatic synonymous mutations altering codon optimality in 24 cancer types (A) Classification of synonymous mutations based on tissue-specific optimal and non-optimal codons. (B) Statistics on the type of alterations in codon optimality. (C) Statistics on the impact of synonymous mutations on codon optimality for synonymous mutations causing one type of codon optimality change. (D) Synonymous mutations (occurrence ≥ 6) causing two different types of alterations in codon optimality. (E) Synonymous mutations causing three different types of alterations in codon optimality. (F) CSC scores changes in HeLa cells for synonymous mutations causing optimal codon gains, optimal codon losses, or no changes in CESC. (G) CSC score changes in HeLa cells for the top 3 most frequent optimal codon gains or losses in CESC. (H) Hierarchical clustering based on the top 3 most frequent optimal codon losses in each cancer type. (I) Hierarchical clustering based on the top 3 optimal codon gains in each cancer type.

Intriguingly, optimal codon gains showed higher proportion in the first 20% of the CDS compared to optimal codon loss (p = 4.7e−05, RR test; Fig. 4B), and were enriched in the 10%-40% region of the CDS compared with no change (p = 0.0004, RR test; Fig. S5A). In addition, optimal codon losses were underrepresented in the first 10% of the CDS compared with no change (p = 6.7e−09, RR test; Fig. S5B), while optimal codon gains were depleted in the 60%-80% region of the CDS when compared with either optimal codon loss (p = 8.1e−06, RR test; Fig. 4B) or no change (p = 0.001, RR test; Fig. S5A). These data suggest that the increase of optimal codons by somatic synonymous mutations preferentially occurs in the coding region near the initiating methionine.

Next, we identified the genes in which optimal codon gains or losses frequently occurred. As a result, synonymous mutations were found to frequently increase optimal codons in 110 genes (Fig. 4C). GO enrichment analysis revealed that these genes were mainly enriched in cell cycle-related processes (Fig. 4D). By contrast, synonymous mutations did not frequently decrease optimal codons in the corresponding mutated genes (Fig. S6A). We also examined the frequency of optimal codon gain or loss in five cancer-related gene sets (i.e., apoptosis, cell cycle, DNA repair, immune response, and metabolism) versus genes nonessential for tumor cell growth. Interestingly, compared with nonessential genes, synonymous mutations were found to frequently increase optimal codons in genes related to DNA repair (p < 2.2e−16, RR test) and cell cycle (p = 3.3e−08, RR test; Fig. 4E). Similar trends were also observed when comparing the frequency of optimal codon gain in individual genes of DNA repair (p = 7e−05, Wilcoxon

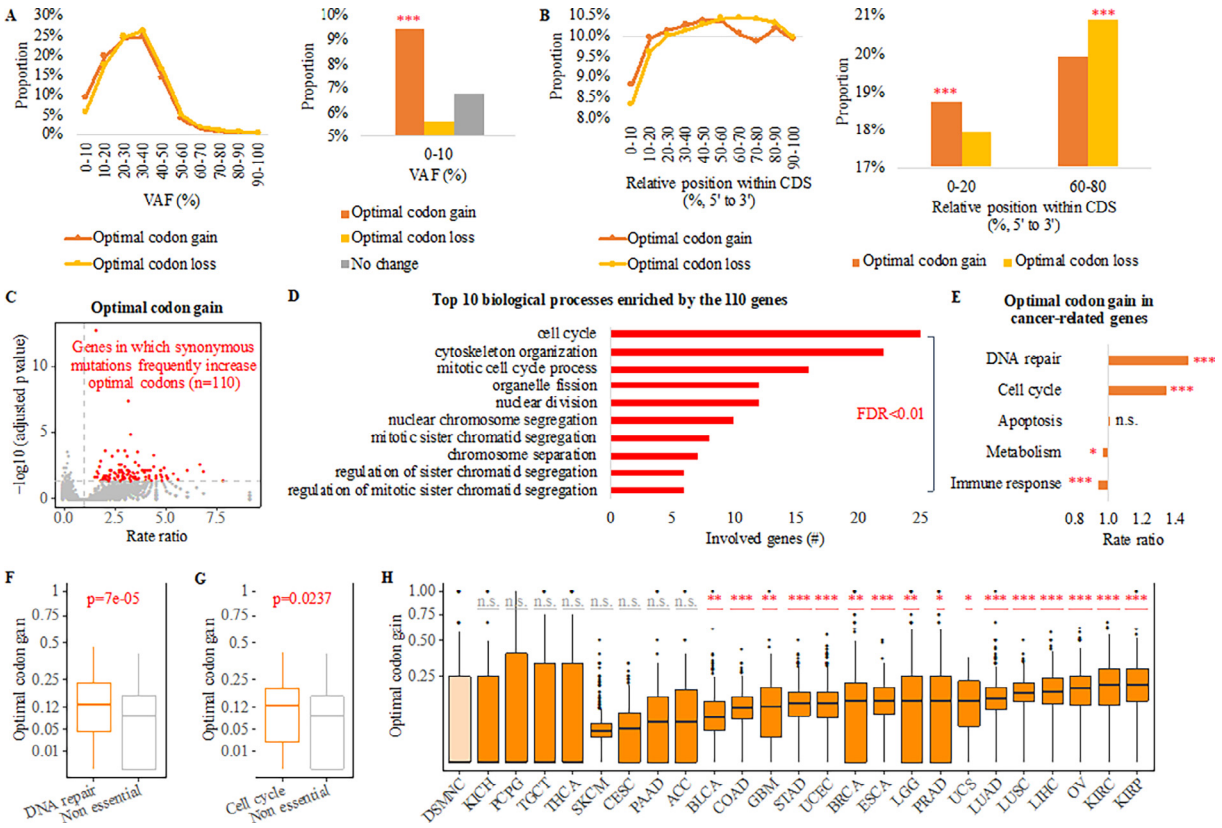

Fig. 4. Synonymous mutations frequently increase optimal codons in 16 cancer types (A) Left: The distribution of VAFs for optimal codon gains or losses. Right: Boxplot showing the distribution of optimal codon gains, optimal codon losses, and no changes in synonymous mutations with VAF <10%. (B) Left: The distribution of optimal codon gains or losses within the CDS. Right: Boxplot showing the distribution of optimal codon gains or losses in the first 20% or 60%–80% of the CDS. (C) Scatter plot showing the genes in which synonymous mutations frequently increase optimal codons (point in red). (D) Top ten biological processes enriched by the genes where synonymous mutations frequently increase optimal codons. (E) The frequency of optimal codon gain in cancer-related genes versus nonessential genes. (F) Boxplot showing the frequency of optimal codon gain in genes related to cell cycle and nonessential genes. (G) Boxplot showing the frequency of optimal codon gain in genes related to DNA repair and nonessential genes. (H) The frequency of optimal codon gain in each cancer types sorted by the median value in each cancer type, compared to the frequency of optimal codon gain in normal cells from DSMNC. n.s., not significant; *, p < 0.05; **, p < 0.01; ***, p < 0.001. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rank-sum test; Fig. 4F) or cell cycle (p = 0.0237, Wilcoxon rank-sum test; Fig. 4G) versus those of nonessential genes. By contrast, although synonymous mutations frequently decreased optimal codons in metabolism-related genes (p = 0.0097, RR test; Fig. S6B), this trend was not observed when comparing the frequency of optimal codon loss in individual metabolism genes versus those of nonessential genes (p = 0.1015, Wilcoxon rank-sum test; Fig. S6C). These data suggest that only the increase of optimal codons by synonymous mutations frequently occurs in cancer-related genes.

In view of the frequent increase of optimal codons by synonymous mutations in cancer-related genes, we next determined whether synonymous mutations increased optimal codons more frequently in tumor cells than in normal cells. For this purpose, we analyzed the impact of synonymous mutations on codon optimality in 416 normal cells from 234 individuals across six human tissues from DSMNC [32], and compared the frequency of optimal codon gain or loss in tumor samples from TCGA with that in normal cells from DSMNC [32]. Interestingly, two-thirds (16/24) of the analyzed cancer types displayed significantly more frequent optimal codon gain compared to normal cells (Fig. 4H). By contrast, only three cancer types (SKCM, CESC, and THCA) exhibited significantly more frequent optimal codon loss than normal cells (Fig. S6D). For validation, somatic mutations in 1,059 normal samples across 28 human tissues from Yizhak et al. [49] were also used for comparison, and similar trends were observed (Fig. S7). Overall, these data suggest that synonymous mutations frequently increase optimal codons in tumor cells and cancer-related genes.

3.5. Elevated frequency of optimal codon gain promotes tumor cell proliferation in three cancer types characterized by DDR deficiency

We next sought to investigate in the above 16 cancer types which cancer hallmarks were associated with the frequent increase of optimal codons by synonymous mutations in tumor cells and cancer-related genes. As the conversion of non-optimal codons to optimal codons was previously demonstrated to increase mRNA stability [9] and expression levels [14], we first identified the genes whose mRNA expression levels increased with an elevated frequency of optimal codon gain. There were 1,390, 1,074, 389, 278, and 3 genes in BRCA, COAD, STAD, UCEC, and KIRP, respectively, whose mRNA expression levels increased with the frequency of optimal codon gain (adjusted p-value <0.05; Fig. S8A). GO enrichment analysis revealed that the identified genes in COAD were enriched in immune-related processes (Fig. S8B). Interestingly, corresponding genes in BRCA, STAD and UCEC were commonly enriched in cell cycle-related processes (Fig. 5A), which pointed to a potential relationship between an elevated frequency of optimal codon gain and increased tumor cell proliferation in these cancer types. Additionally, those genes in BRCA were also enriched in DNA repair and cellular response to DNA damage stimulus (Fig. 5A).

To determine whether an elevated frequency of optimal codon gain promoted tumor cell proliferation, two indices reflecting tumor cell proliferation (i.e., ki-67 mRNA expression level, and proliferation from Thorsson et al. [51]) were employed to assess the association between the frequency of optimal codon gain and tumor cell proliferation. As a result, significantly positive correlations were observed between the frequency of optimal codon gain and both indices in three cancer types including BRCA, STAD, and UCEC (Fig. 5B-C). Of note, the three cancer types were exactly the cancer types where cell cycle-related processes were enriched by the genes whose mRNA expression levels increased with elevated frequency of optimal codon gain, demonstrating that elevated frequency of optimal codon gain promoted tumor cell proliferation in these cancer types. In addition, the frequency of optimal codon gain was positively associated with ki-67 mRNA expression level in PRAD (rho = 0.12, adjusted p = 0.0448; Fig. 5B).

Notably, HR deficiency frequently occurs in BRCA [57], while STAD and UCEC are associated with MMR deficiency [58,59]. We hypothesized that DDR deficiency may be involved in elevated frequency of optimal codon gain in these cancer types. To test this hypothesis, we quantified eight types of DDR deficiency (i.e., BER, DR, FA, HR, MMR, NER, NHEJ, and TLS) based on the genomic alterations in 71 DNA repair pathway-specific genes from Knijnenburg et al. [52] and analyzed their associations with the frequency of optimal codon gain in BRCA, STAD, and UCEC. In all the three cancer types, significantly positive correlations were observed between the frequency of optimal codon gain and five types of DDR deficiency: BER, HR, MMR, NER, and NHEJ (Fig. 5D). Additionally, an elevated frequency of optimal codon gain correlated with increased DR deficiency in BRCA and STAD, and with enhanced FA deficiency in STAD and UCEC. We further examined the relationship between the frequency of optimal codon gain and HR defects obtained from Thorsson et al. [51]. In line with aforementioned results, significant positive correlations were observed between the frequency of optimal codon gain and HR defects in all the three cancer types (BRCA: rho = 0.22, p = 2.3e−11; STAD: rho = 0.15, p = 0.0021; UCEC: rho = 0.10, p = 0.0283; Fig. 5E). These data suggest that DDR deficiency may be related to an elevated frequency of optimal codon gain in these cancer types, which may in turn promote tumor cell proliferation.

3.6. Elevated frequency of optimal codon gain correlates with better survival in patients with TNBC

Notably, DDR competency is a determinant of sensitivity to platinum-based compounds used for cancer chemotherapy, which exert their cytotoxic effects by inducing DNA damage [60]. Therefore, we determined whether the frequency of optimal codon gain was of prognostic significance in the above cancer types (i.e., BRCA, STAD, and UCEC). We found that decreased frequency of optimal codon gain was significantly correlated with better overall survival (OS) in patients with BRCA (hazard ratio, 0.66; 95% CI, 0.45–0.96; p = 0.029) as well as with improved progression-free survival in patients with UCEC (hazard ratio, 0.68; 95% CI, 0.47–0.98; p = 0.037; Fig. S9A). After adjusting for clinicopathological parameters, the association between decreased frequency of optimal codon gain and better OS remained statistically significant in BRCA (hazard ratio, 0.65; 95% CI, 0.44–0.95; p = 0.028; Fig. S9B). This suggests that the frequency of optimal codon gain is a potential prognostic biomarker for patients with BRCA.

Since BRCA can be classified as luminal A, luminal B, HER2+, or TNBC [54], we next determined whether the association between the frequency of optimal codon gain and patient survival differed between these subtypes. Interestingly, decreased frequency of optimal codon gain trended to correlate with better survival in patients with luminal A (hazard ratio, 0.55; 95% CI, 0.26–1.18; p = 0.12), luminal B (hazard ratio, 0.58; 95% CI, 0.33–1.04; p = 0.065), or HER2 + BRCA (hazard ratio, 0.21; 95% CI, 0.02–

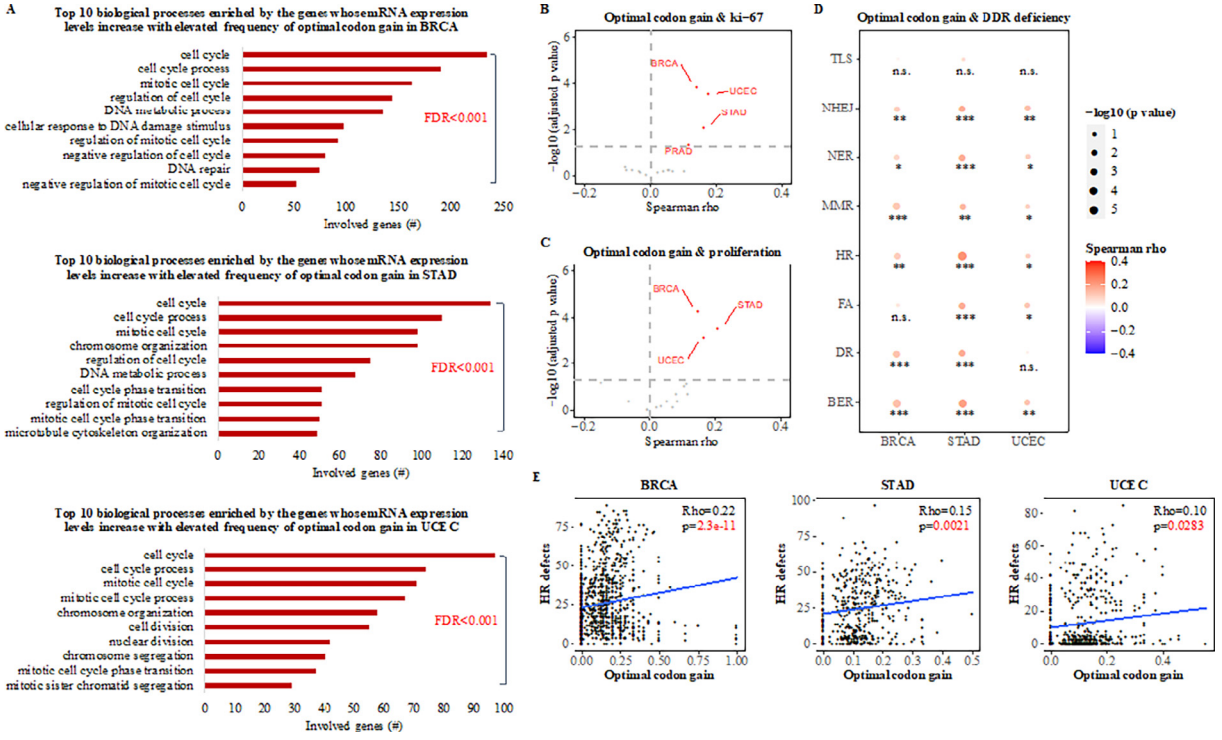

Fig. 5. Elevated frequency of optimal codon gain promotes tumor cell proliferation in three cancer types characterized by DDR deficiency (A) Top 10 biological processes enriched by the genes whose mRNA expression levels increase with the frequency of optimal codon gain in BRCA, STAD, and UCEC. (B-E) The association between the frequency of optimal codon gain and ki-67 mRNA expression levels (B), proliferation (C), eight types of DDR deficiency (D), or HR defects obtained from Thorsson et al. [51] (E). n.s., not significant; *, p < 0.05; **, p < 0.01; ***, p < 0.001.

1.92; p = 0.13; Fig. S10). Conversely, decreased frequency of optimal codon gain was significantly associated with worse outcome in patients with TNBC (hazard ratio, 6.05; 95% CI, 1.25–29.18; p = 0.014; Fig. 6A), and this association remained statistically significant after adjusting for clinicopathological parameters (hazard ratio, 14; 95% CI, 1.61–122.0; p = 0.017; Fig. 6B). To confirm the prognostic role of the frequency of optimal codon gain in patients with TNBC, we analyzed the impact of somatic synonymous mutations on codon optimality for 147 patients with TNBC receiving adjuvant chemotherapy from Staaf et al. [55]. In line with the above results, tumors with a decreased frequency of optimal codon gain trended to have a worse distant relapse-free interval (DRFI; hazard ratio, 2.20; 95% CI, 0.92–5.26; p = 0.069; Fig. 6C) and had a significantly worse invasive disease-free survival (IDFS; hazard ratio, 2.28; 95% CI, 1.14–4.54; p = 0.016; Fig. 6C) compared with tumors with an elevated frequency of optimal codon gain. Additionally, statistical significance remained after adjusting available clinicopathological parameters (hazard ratio, 2.13; 95% CI, 1.05–4.3; p = 0.036; Fig. 6D). Overall, these data suggest that an elevated frequency of optimal codon gain is an independent prognostic biomarker in patients with TNBC.

4. Discussion

Synonymous mutations have long been assumed to be neutral and provide no fitness advantage to tumor cells. With the availability of large-scale cancer genomics data, synonymous mutation was recently shown to be functional in human cancers for the past few years via altering motif regulating splicing [24], mRNA secondary structure [4], RNA binding proteins, and miRNA binding sites [61]. However, the impact of synonymous mutation on codon optimality remains largely unexplored in human cancers. Here, we generated tissue-specific codon optimality in 29 human tissues and employed this data to investigate how codon optimality was perturbed by synonymous mutations in human cancers.

An effort was previously made by Supek et al. [24] to examine the impact of synonymous mutations on codon optimality in human cancers, yet codon optimality was determined using the genomic copy number of tRNA genes and the frequency of optimal codon gain was only examined in 16 known oncogenes highly enriched with synonymous mutations. These might be the reasons why synonymous mutations were not found to frequently increase optimal codons in human cancers. Therefore, a key step of our study was to generate tissue-specific codon optimality for human tissues. Since optimal codons are thought to be translated faster than non-optimal codons by the ribosome [23], codon-specific ribosome density, which could reflect translational efficiency to some extent, was taken as one of the golden standards in the evaluation of which metric measured codon optimality more accurately in *Homo sapiens*. Additionally, mRNA stability-based CSC score, served as the other golden standard in the evaluation. That C_{opt} correlated positively with CSC score and negatively with

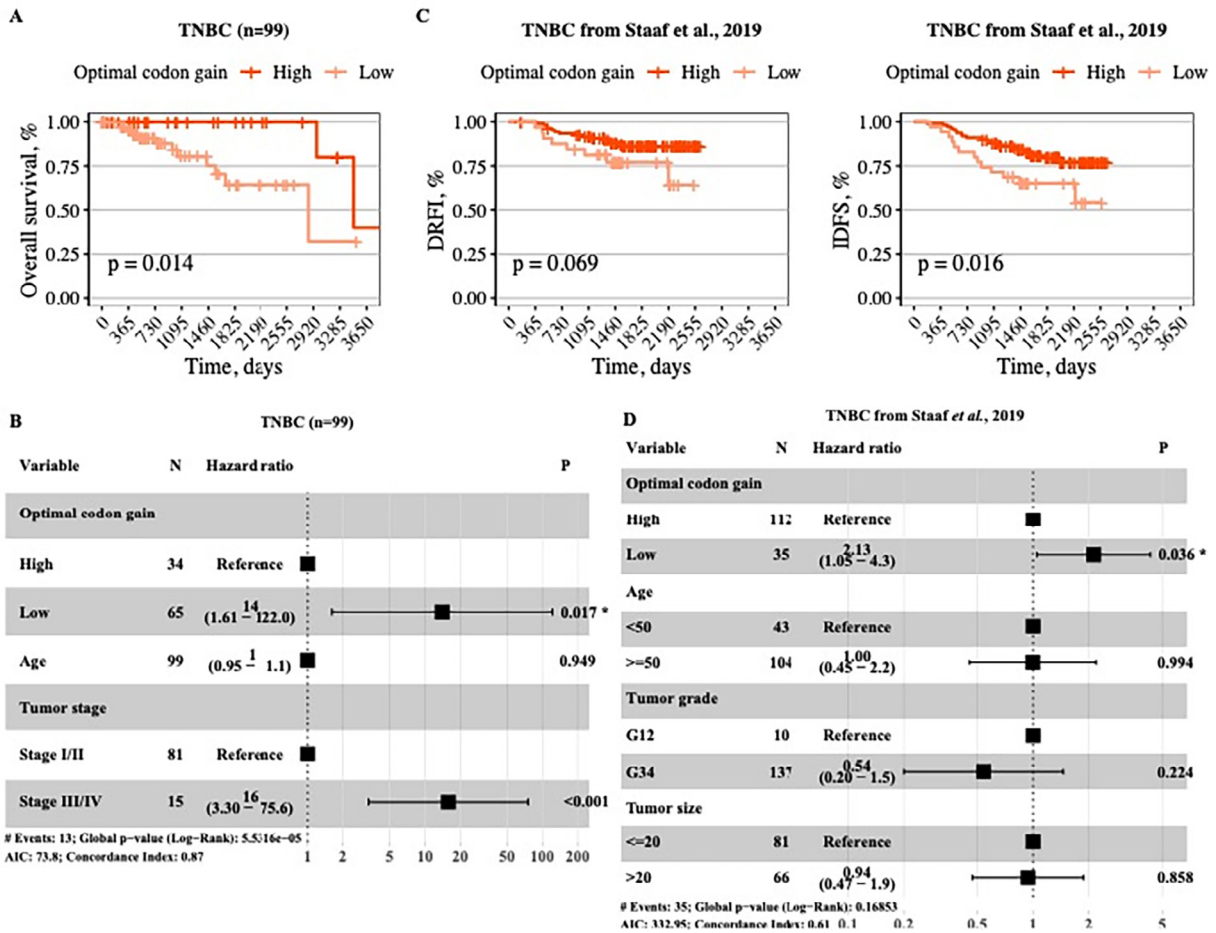

Fig. 6. Elevated frequency of optimal codon gain correlates with better survival in patients with TNBC (A) Association between the frequency of optimal codon gain and overall survival in patients with TNBC from TCGA. (B) Multivariable analysis of the frequency of optimal codon gains and available clinicopathologic parameters (i.e., age at diagnosis and tumor stage) in patients with TNBC from TCGA. (C) Association of the frequency of optimal codon gain with DRFI and IDFS in patients with TNBC receiving adjuvant chemotherapy from Staaf et al. [55]. (D) Multivariable analysis of the frequency of optimal codon gain and available clinicopathologic parameters (i.e., age at diagnosis, tumor grade, and tumor size) in patients with TNBC from Staaf et al. [55].

codon-specific ribosome density simultaneously demonstrated that C_{opt} could be employed to generate tissue-specific codon optimality for human tissues. Furthermore, RR score, our modified version of C_{opt}, showed stronger correlations with both CSC score and codon-specific ribosome density than C_{opt}, suggesting that RR score is a better measurement of codon optimality in *Homo sapiens*. Employing RR score to generate tissue-specific codon optimality in 29 human tissues, we found that optimal and non-optimal codons generally ended with C/G and A/T, respectively. This was in agreement with recent findings that codons ended with C/G stabilized mRNA while those ended with A/T destabilized mRNA [9,12]. Furthermore, we found that the tissue specificity of codon optimality was primarily present in amino acids with high genetic code degeneracy. This is in line with the biological significance of degeneracy in genetic code – reducing the probability of an error caused by the mutational substitution of a base in the triplet, which makes it possible for organisms to survive and prosper [62].

Applying tissue-specific codon optimality to somatic synonymous mutations in tumor samples across 24 cancer types from TCGA, we found that 62% of synonymous mutations caused optimal codon losses, and only 12 percent led to optimal codon gains. The high frequency of optimal codon loss implies that, tumor progression is unlikely to be promoted by the loss of optimal codons, since harmful mutations will be eliminated by purifying selection before they reach high frequency in the population [63]. By the same token, the low frequency of optimal codon gain, alongside with the preferential increase of optimal codons by synonymous mutations with VAF $\leq$10%, indicates that synonymous mutations may contribute to tumor progression via increasing optimal codons. In support of this, synonymous mutations were found to frequently increase optimal codons in 16 cancer types, yet in only three cancer types did synonymous mutations frequently decrease optimal codons. Furthermore, only the increase of optimal codons was found to frequently occur in cancer-related genes (i.e., cell cycle and DNA repair), which was in line with recent findings that genes involved in DNA damage response and cell cycle were particularly intolerant to synonymous mutations [25]. Although the frequency of optimal codon gain was found to be increased in genes related to both cell cycle and DNA repair when analyzed as a gene set, genes in which synonymous mutations frequently increased optimal codons were mainly enriched in cell cycle-related processes. This suggests that cell cycle-related genes are more heavily influenced by the increase of optimal codons by synonymous mutations. This may be due to the association of non-optimal codons with proliferation, which indicates that the frequent substitution of non-optimal with optimal codons in tumor cells may largely occur in genes related to proliferation (e.g., cell cycle). In support of this, cell cycle-related processes were enriched by genes whose mRNA expression levels increased with elevated frequency of optimal codon gain in all the three cancer types (i.e., BRCA, STAD, and UCEC) where elevated frequency of optimal codon gain promoted tumor cell proliferation, while DNA repair-related processes were only enriched by corresponding genes in BRCA. Overall, these data suggest that synonymous mutations frequently increase optimal codons in tumor cells and cell cycle-related genes.

We uncovered a positive relationship between the frequency of optimal codon gain and tumor cell proliferation in three cancer types including BRCA, STAD and UCEC. Notably, BRCA is one of the cancer types where HR deficiency most frequently occurs [57], while STAD and UCEC are among the cancer types in which MMR deficiency frequently occurs [58,59]. This indicates that DDR deficiency may be involved in elevated frequency of optimal codon gain in tumor cells and cancer-related genes. In support of this, a positive relationship was uncovered between the frequency of optimal codon gain and DDR deficiency in the above cancer types. Furthermore, in line with previous observations that HR deficiency correlated with increased response to DNA-damaging platinum-containing therapy in patients with TNBC [64], elevated frequency of optimal codon gain was found to be associated with improved outcome in patients with TNBC from TCGA and those receiving adjuvant chemotherapy from an independent cohort. These data suggest that the elevated frequency of optimal codon gain may be a footprint of DDR deficiency that has long been neglected. It should be noted, however, that there may be other genomic or epigenomic events in addition to DDR deficiency resulting in an elevated frequency of optimal codon gain in tumor cells, as the promotion of tumor cell proliferation by the elevated frequency of optimal codon gain was not observed in OV where HR deficiency also frequently occurs [57].

While our analyses suggest that differences in mutation rate won't have much influence on the frequency of optimal codon gain or loss (Fig. S11), we cannot completely rule out the influence of mutation rate differences on the frequency of optimal codon gain or loss. In addition, although a direct comparison was not conducted between expression data from GTEx and TCGA, we cannot entirely exclude the possibility that some observations may be slightly influenced by differences between these two databases. Nevertheless, such differences may not affect the main results of this study, as consistent results were observed from different analyses. For example, genes whose expression levels increased with an elevated frequency of optimal codon gain were mainly enriched in processes related to cell cycle in BRCA, STAD and UCEC, which was consistent with the observation that genes in which synonymous mutations frequently increase optimal codons were mainly enriched in processes related to cell cycle. In support of these observations, the frequency of optimal codon gain showed positive association with proliferation from Thorsson et al. [51] and ki-67 mRNA expression levels in BRCA, STAD and UCEC.

In summary, we have profiled tissue-specific codon optimality in human tissues and uncovered alterations in codon optimality in human cancers. The tissue-specific codon optimality determined in this study may be used to improve the translation efficiency of recombinant genes in developing personalized mRNA vaccines for infectious diseases or cancers. A better understanding of the alterations in codon optimality in human cancers will undoubtedly provide insights into the development of novel cancer therapies. Future studies may be conducted to answer whether missense mutations contribute to tumor progression and outcome via altering codon optimality in addition to altering the amino acid composition of the encoded proteins.

CRediT authorship contribution statement

Xia Ran: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Supervision, Project administration. **Jinyuan Xiao:** Data curation, Software, Formal analysis. **Fang Cheng:** Visualization, Formal analysis. **Tao Wang:** Writing – review & editing. **Huajing Teng:** Writing – review & editing. **Zhongsheng Sun:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Zhipeng Zhou and Wenfeng Qian for constructive suggestions and critical reading of the manuscript. We

would like to thank Jinyang Zhang and Fangqing Zhao for helpful advices.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32170650).

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.005.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(3):209–49.
[2] Garrido-Castro AC, Lin NU, Polyak K. Insights into molecular classifications of triple-negative breast cancer: improving patient selection for treatment. Cancer Discovery 2019;9(2):176–98.
[3] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47(D1):D941–7.
[4] Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Gross M, et al. A pan-cancer analysis of synonymous mutations. Nat Commun 2019;10(1):2569.
[5] Quax TEF, Claassens NJ, Soll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell 2015;59(2):149–61.
[6] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 2011;12(1):32–42.
[7] Liu Y, Yang Q, Zhao F. Synonymous but not silent: the codon usage code for gene expression and protein folding. Annu Rev Biochem 2021;90:375–401.
[8] Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol 2018;19(1):20–30.
[9] Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell 2015;160(6):1111–24.
[10] Burow DA, Martin S, Quail JF, Alhusaini N, Coller J, Cleary MD. Attenuated codon optimality contributes to neural-specific mRNA decay in Drosophila. Cell Rep 2018;24(7):1704–12.
[11] Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, et al. Translation affects mRNA stability in a codon-dependent manner in human cells. Elife 2019;8.
[12] Hia F, Yang SF, Shichino Y, Yoshinaga M, Murakawa Y, Vandenbon A, et al. Codon bias confers stability to human mRNAs. EMBO Rep 2019;20(11):e48220.
[13] Medina-Munoz SG, Kushawah G, Castellano LA, Diez M, DeVore ML, Salazar MJB, et al. Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. Genome Biol 2021;22(1):14.
[14] Chen S, Li K, Cao W, Wang J, Zhao T, Huan Q, et al. Codon-resolution analysis reveals a direct and context-dependent impact of individual synonymous mutations on mRNA level. Mol Biol Evol 2017;34(11):2944–58.
[15] Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A 2016;113(41):E6117–25.
[16] Meyer D, Kames J, Bar H, Komar AA, Alexaki A, Ibla J, et al. Distinct signatures of codon and codon pair usage in 32 primary tumor types in the novel database CancerCoCoPUTs for cancer-specific codon usage. Genome Med 2021;13(1):122.
[17] Benisty H, Weber M, Hernandez-Alias X, Schaefer MH, Serrano L. Mutation bias within oncogene families is related to proliferation-specific codon usage. Proc Natl Acad Sci U S A 2020;117(48):30848–56.
[18] Uddin A, Paul N, Chakraborty S. The codon usage pattern of genes involved in ovarian cancer. Ann N Y Acad Sci 2019;1440(1):67–78.
[19] Chakraborty S, Paul S, Nath D, Choudhury Y, Ahn Y, Cho YS, et al. Synonymous codon usage and context analysis of genes associated with pancreatic cancer. Mutat Res 2020;821:111719.
[20] McCarthy C, Carrea A, Diambra L. Bicodon bias can determine the role of synonymous SNPs in human diseases. BMC Genomics 2017;18(1):227.
[21] dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 2004;32(17):5036–44.
[22] Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol 2009;26(7):1571–80.
[23] Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 2013;20(2):237–43.
[24] Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell 2014;156(6):1324–35.
[25] Dhindsa RS, Copeland BR, Mustoe AM, Goldstein DB. Natural selection shapes codon usage in the human genome. Am J Hum Genet 2020;107(1):83–95.
[26] Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA expression. PLoS Genet 2006;2(12):e221.
[27] Huang J, Chen W, Zhou F, Pang Z, Wang L, Pan T, et al. Tissue-specific reprogramming of host tRNA transcriptome by the microbiome. Genome Res 2021;31(6):947–57.
[28] Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci USA 2004;101(34):12588–91.
[29] Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, et al. TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. J Mol Biol 2020;432(11):3369–78.
[30] Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L. Translational efficiency across healthy and tumor tissues is proliferation-related. Mol Syst Biol 2020;16(3):e9275.
[31] Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348(6235):648–60.
[32] Miao X, Li X, Wang L, Zheng C, Cai J. DSMNC: a database of somatic mutations in normal cells. Nucleic Acids Res 2019;47(D1):D971–5.
[33] Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 2016;44(D1):D184–9.
[34] Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, et al. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods 2015;12(9):835–7.
[35] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20.
[36] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 2011;12:323.
[37] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
[38] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 2012;7(8):1534–50.
[39] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011.
[40] Liu Q, Shvarts T, Sliz P, Gregory RI. RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. Nucleic Acids Res 2020;48(W1):W218–29.
[41] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10(3):R25.
[42] Li F, Xing X, Xiao Z, Xu G, Yang X. RiboMiner: a toolset for mining multi-dimensional features of the translatome with ribosome profiling data. BMC Bioinf 2020;21(1):340.
[43] Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep 2016;14(7):1787–99.
[44] Shi MW, Zhang NA, Shi CP, Liu CJ, Luo ZH, Wang DY, et al. SAGD: a comprehensive sex-associated gene database from transcriptomes. Nucleic Acids Res 2019;47(D1):D835–40.
[45] Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, et al. A dual program for translation regulation in cellular proliferation and differentiation. Cell 2014;158(6):1281–92.
[46] Ran X, Xiao J, Zhang Y, Teng H, Cheng F, Chen H, et al. Low intratumor heterogeneity correlates with increased response to PD-1 blockade in renal cell carcinoma. Ther Adv Med Oncol 2020;12:1758835920977117.
[47] Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, et al. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. Cell 2020;180(5):915–27 e16.
[48] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol 2016;17(1):122.
[49] Yizhak K, Aguet F, Kim J, Hess JM, Kubler K, Grimsby J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science 2019;364(6444).
[50] Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res 2019;47(W1):W199–205.
[51] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. Immunity 2018;48(4):812–810 e14.
[52] Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome Atlas. Cell Rep 2018;23(1):239–54 e6.

[53] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173(2):400–16 e11.
[54] Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 2013;24(9):2206–23.
[55] Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reutersward C, Hakkinen J, et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med 2019;25(10):1526–33.
[56] Fornasiero EF, Rizzoli SO. Pathological changes are associated with shifts in the employment of synonymous codons at the transcriptome level. BMC Genomics 2019;20(1):566.
[57] Nguyen L, J WMM, Van Hoeck A, Cuppen E.. Pan-cancer landscape of homologous recombination deficiency. Nat Commun 2020;11(1):5584.**.
[58] van Velzen MJM, Derks S, van Grieken NCT, Haj Mohammad N, van Laarhoven HWM. MSI as a predictive factor for treatment outcome of gastroesophageal adenocarcinoma. Cancer Treat Rev 2020;86:102024.
[59] Stelloo E, Jansen AML, Osse EM, Nout RA, Creutzberg CL, Ruano D, et al. Practical guidance for mismatch repair-deficiency testing in endometrial cancer. Ann Oncol 2017;28(1):96–102.
[60] Birkbak NJ, Wang ZC, Kim JY, Eklund AC, Li Q, Tian R, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov 2012;2(4):366–75.
[61] Teng H, Wei W, Li Q, Xue M, Shi X, Li X, et al. Prevalence and architecture of posttranscriptionally impaired synonymous mutations in 8,320 genomes across 22 cancer types. Nucleic Acids Res 2020;48(3):1192–205.
[62] Ikehara K. Degeneracy of the genetic code has played an important role in evolution of organisms. international journal of genetic. Science 2016.
[63] Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. Cell 2019;177(1):70–84.
[64] Telli ML, Timms KM, Reid J, Hennessy B, Mills GB, Jensen KC, et al. Homologous Recombination Deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast Cancer. Clin Cancer Res 2016;22(15):3764–73.